基于异构数据库的空间天文卫星数据组织方法

杨晓艳[1,2]，孙小涓[1,2,3]，石涛[1,2]，刘志奇[1]，佟继周[4]

（1.中国科学院空天信息创新研究院，北京 100190；2.中国科学院空间信息与应用系统重点实验室，北京 100190；3.中国科学院大学，北京 100049；4.中国科学院国家空间科学中心，北京 100190）

摘 要：随着空间天文卫星获取数据量越来越多，数据应用逐渐发挥了重要作用。现有的天文卫星地面系统中，数据存储方式和组织方法各异，数据量达 PB 级，并且数据量持续增长，无法快速查找并提取特征参数，难以满足数据应用对查询的时效性要求。本文提出了一种新的空间天文卫星数据组织方法，通过解析抽取数据中的海量特征参数，建立观测时间、空间位置与特征参数的关联，实现在统一时空下的多源数据组织；同时采用关系型数据库与非关系型数据库结合的异构存储方式，设计了海量特征参数存储管理系统。将本文方法应用于空间科学卫星大数据应用平台系统中，使用硬 X 射线调制望远镜卫星数据的实验结果表明，系统针对按照时间、空间条件获取数据的要求能够较好满足，相比关系型数据库数据组织方式，相同查询模式下数据检索效率明显提高；并且随着数据存储量的增加，系统具有稳定的扩展能力。

关键词：空间天文数据；时空索引；数据组织；非关系型数据库

中图分类号：P171.3

2015 年以来，我国陆续发射了暗物质粒子探测卫星、硬 X 射线调制望远镜、引力波暴高能电磁对应体全天监测器卫星等空间天文卫星，持续获取了大量空间天文观测数据。地面系统对卫星原始探测数据以及在此基础上生成的编辑级产品、标定级产品进行存储和管理，这些数据产品是卫星在一定时空条件下的探测成果，产品内容包含粒子类别、粒子数量、粒子能段、粒子入射径迹、粒子能量沉积等表征空间天文目标的信息，同时还包含卫星在轨姿态、轨道位置、温度、压力等表征卫星平台、有效载荷工作状况的信息。这些数据用于空间天文研究、卫星及载荷健康状况趋势分析、卫星探测目标分析与计划辅助制定、卫星探测过程可视化等应用领域，能够发挥重要的数据价值。

现有的空间科学战略先导专项卫星数据地面管理系统中，按照国家空间科学中心提出空间科学数据模型[1]，空间天文卫星数据以 FITS（Flexible Image Transport System）[2-3]、ROOT（欧洲核子研究中心开发的一种数据格式）等空间科学领域专用的数据格式保存在文件中。在获取数据时，首先需要检索数据文件并解析文件格式[4-5]，然后从文件指定位置抽取所需的特征参数，对某些数据还需要进行物理量转换、时间校正等处理[6]。由于各型卫星数据产品的存储格式不相同，获取特征参数的数据处理过程也不相同，处理过程复杂且耗时，而随着数据量的不断增长，数据库检索时间越来越长，数据获取的实时性越来越难以保证。目前以文件为粒度进行数据存储和组织的系统难以满足数据实时检索获取的要求。

为了满足实时获取数据的应用需求，需要从空间天文数据文件中抽取出特征参数，构建一种高效的参数级细粒度数据组织方法。但是，从海量空间天文数据文件中抽取得到的特征参数数量巨大，如何高效地组织和索引这些数据将是一个非常关键的问题。

1 空间天文卫星数据特点

空间天文卫星观测对象主要是宇宙太空中的各类天体目标，空间天文卫星数据包括表征这些观测对象的科学数据，以及表征卫星和载荷状态的工程数据，这类数据存在以下特点：

（1）数据种类多样，时间分辨率高，数据量庞大

从产品内容来讲，空间天文卫星数据包含天文目标科学数据、卫星平台及载荷工程数据等类型；从产品级别来讲，包含编辑级产品、标定级产品等类型。每颗卫星的产品内容、产品级别有所不同。以暗物质卫星为例，产品级别有 9 级，每级产品类型约为十几类，共计 100 多类。以暗物质卫星标定级产品为例，半小时左右的数据文件中包含粒子数量达 12 万左右，每个粒子的参数包括粒子在各载荷中的沉积能量、击中位置、粒子入射径迹等，按 5 年卫星寿命期估算，产生的数据约为 105.1 亿条。工程数据包括卫星 AOCC（Attitude and Orbit Control Computer）姿态数据、GPS（Global Positioning System）定位数据等几十类数据，大部分数据为每秒一条记录，还有一部分数据每秒两条甚至四条记录。按照每秒一条记录进行估算，每颗卫星每年每类数据产生 3000 多万条记录，按照卫星寿命 5 年、每颗卫星 35 类数据估算，单星寿命期内产生数据量约 50 多亿条。数据总量达百亿甚至千亿数量级，迫切需要构建一种针对海量多源数据的高效组织方法。

资助项目：中国科学院战略性先导科技专项空间科学（二期）科学卫星地面支撑项目；空间科学卫星大数据应用承载平台项目

作者简介：杨晓艳，女，硕士，研究方向为空间科学卫星数据处理与组织。Email: yangxiaoyan@mail.ie.ac.cn

（2）数据具有时间、空间属性特征，需要支持基于时空条件快速检索多类数据的应用需求

空间天文卫星数据信息表达为（Time，RA，DEC，par1，par2……）。其中，Time 表示观测时间；RA 表示当前观测时间卫星视场中心点赤经；DEC 表示当前观测时间卫星视场中心点赤纬；par1，par2 表示特征参数值，比如高能电子计数值、载荷工程参数测量值等。空间天文数据具有时间、空间属性特征，为了支持基于时空检索条件对多源数据进行快速检索，需要对多源数据的时间、空间属性进行统一处理，构建基于特征参数时空索引的数据组织方法和检索方法，面向应用提供符合要求的数据。

（3）数据量持续增长，需要可扩展架构支持日益增长的数据存储要求

伴随着已有卫星持续在轨运行和新型卫星发射入轨，空间天文卫星数据体量呈现持续增长的趋势，需要构建一套在存储容量方面具备良好可扩展性的分布式数据库存储系统；并且，随着存储容量的增加，其检索效率能够基本稳定。

2 相关研究工作

面对空间天文卫星海量数据组织和快速获取的需求，传统关系型数据库难以满足。HBase 等非关系型数据库具有数据结构灵活、水平扩展性强的特点，能够比传统的结构化数据库更加有效地组织大数据[12]。然而，由于 HBase 数据库仅在主键上建立了 B+树索引，能够提供基于主键的快速查询能力；在面对非主键查询请求时，需要进行全表扫描，导致查询效率很低。而空间天文卫星数据需要按照时间、赤经、赤纬、参数等多重属性进行检索，HBase 难以满足按照多重属性快速检索数据的需求。

利用非关系型数据库存储和检索海量时空数据，多个行业的学者都进行了研究，主要分为两种思路：一种是地理信息、国土资源、空间科学等领域的学者，从构建时空格网模型出发，将时空数据按照时空编码存入非关系型数据库。比如张嘉等[7]提出了一种空间矢量数据分布式存储模型，采用四叉树建立空间格网，并以格网编号、随机码构建行键，将数据存储在 HBase 数据库中。康栋贺等[8]提出了 HTM-ST 离散化时空数据组织模型，采用时间、空间离散剖分的方式建立时空耦合编码，并以该编码构建行键，将日地空间数据存储在 HBase 数据库中。由于 HBase 采用字典序方式存储行键，采用多重属性构建行键的方法仅适用于点查询；针对范围查询，需要逐层判断各个时空网格与查询范围的拓扑关系，在剖分细化的过程中不断逼近查询条件中的时空范围，或者进行全表扫描，查询耗时明显。

另一种思路是计算机信息技术领域的学者，通过构建多层索引，来提升非关系型数据库的数据检索效率。比如，葛微等[9]提出一种基于索引表和值表、并结合热点数据缓存的分层式索引技术。该方法在一定程度上提升了检索效率，但在多属性范围检索时需要对多列查询结果进行合并处理，同样无法满足空间科学领域需要按照时空范围实时获取数据的需求。袁茂林等[10]提出一种名为 TA-index 的三层索引技术。该方法旨在提升数据入库效率，在时空范围查询时由于需要分多次查询多层索引和数据库表，因此耗时较长。

针对海量空间天文卫星数据需要按照时间、空间双重属性进行组织和查询的需求，本文提出了提出了一种新的空间天文卫星数据组织方法。首先解析数据文件并从中抽取出海量特征参数，建立观测时间、空间位置与特征参数的关联，实现在统一时空下的多源数据组织。然后结合非关系型数据库数据结构灵活、水平扩展性强，以及关系型数据库在多列值范围查询方面的优势，建立了一套基于异构数据库进行数据组织和存储的方式，其中采用分布式数据库分区分表的方式，构建空间天文卫星 HBase 集群数据库，对海量特征参数进行存储管理；采用关系型数据库分表的方式，存储空间天文卫星时空索引数据，支持从时间、空间两个维度检索数据。

3 空间天文卫星数据组织

3.1 特征参数抽取

现有的空间天文卫星数据以文件形态存储在地面管理系统中，特征参数抽取是空间天文卫星大数据高效组织的第一步。基于 FITSIO（https://heasarc.gsfc.nasa.gov/fitsio/fitsio.html）、ROOT（http://root.cern.ch）格式解析框架，构建数据解析算法，能够适应现有卫星多种数据格式参数抽取的要求。主要步骤如下：

（1）预先设置各卫星各类数据需要抽取的参数，生成参数抽取需求；

（2）获取卫星数据产品文件，识别卫星名称、数据类型、存储格式；

（3）根据卫星名称、数据类型，与步骤（1）的参数抽取需求进行匹配；

（4）针对 FITS 格式数据文件，调用 FITSIO 提取各参数值及其观测时间；针对 ROOT 格式数据文件，调用 ROOT 格式解析框架提取各参数值及其观测时间；针对 CSV、dat 等普通格式文件，直接提取各参数

值及其观测时间；

（5）根据需要对提取出的参数进行物理量转换，比如将载荷温度、压力等参数值从星上记录的电信号值转换为有物理含义的参数值；

（6）基于卫星姿态数据计算逐时刻下卫星观测视场位置信息（RA，DEC）。

3.2 特征参数存储

面向海量特征参数时间序列数据的存储需求，本文提出一种基于 HBase 集群构建的特征参数存储结构（如图 1 所示），以参数分表+时间分区的方式进行存储，支持以时间点、时间范围为条件检索数据。

首先，以单个参数或几个关联参数为粒度，将海量特征参数划分为 Group1、Group2 等不同的参数组，分别建立参数表。其中，针对姿态四元数、轨道位置 xyz 坐标值、轨道六根数等互相关联的参数，将其按组建表存储，比如图中参数 A、B、C 划分为一组，参数 U、V 划分为一组；其余参数单独建表存储。这种存储方式一方面能够提高数据存储的灵活性，便于管理；另一方面能够支持多组参数表并发查询，从而提高多参数查询效率。

然后，基于各类参数的时间频率，按照时间范围对各参数表进行分区，分别建立独立的时间分区索引。比如，图 1 中 Table1 参数时频较高，以 5 个时间单位为跨度建立 t1、t6、t11…的时间分区索引；而 TableN 参数时频较低，以 10 个时间单位为跨度建立 t1、t11…的时间分区索引。这种分区设计能够将数量庞大的参数按照时间范围存入不同区域，在参数检索过程中，支持通过分区索引查找对应时段数据，并且能够支持多分区并发查询，从而进一步提高查询效率。

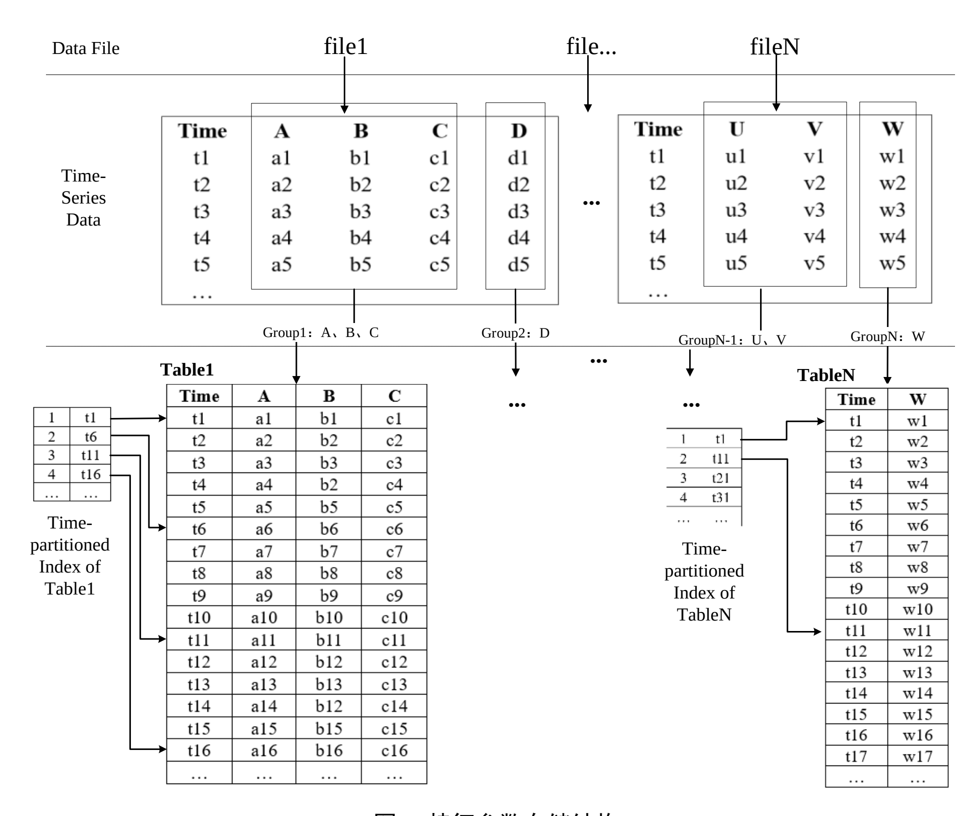


图 1 特征参数存储结构

Fig.1 Parameters' storage structure

3.3 时空索引存储

时空索引表示观测时间与观测视场中心点位置的关系，需要存储 Time、RA、DEC 三个字段数据。时空索引需要满足按照时间、空间范围联合检索的需求，即按照 Time、RA、DEC 字段范围获取数据。由于非关系型数据库 HBase 的优势在于通过行键或者行键的范围快速检索数据，在面对非主键查询需求时，需要进行全表扫描，因而效率较低。而关系型数据库采用 SQL（Structured Query Language，结构化查询语言）查询的方式，适合这种多列值查询的应用需求，不仅能够满足点查询需求（按照指定时间、位置获取数据），而且能够满足范围查询需求（按照时间范围、空间范围获取数据）。因此，本文将时空索引数据存入关系型数据库 MySQL 库表中。

卫星观测过程中，每秒产生一条时空索引数据，观测时间为顺序递增值。在时空索引表中，将 Time 字段设置为主键。另外，由于时空索引每秒一条记录，每颗卫星每年数据量高达 3000 多万条，而 MySQL

库表数据到了千万级后，检索效率会很低。对数据库表进行水平切分，能够解决超大型数据量和高负载库表遇到瓶颈的问题，提高检索效率。

由于本文面向的典型应用场景每次请求数据的时长基本上是小时级，大概率是查询单表，而单表数据量控制在百万级别能够保证检索效率。因此，在时空索引数据存储过程中，按照观测时间 Time 字段，以月为单位对时空索引表进行水平切分，切分后的子表数据量为 200 多万条。同时，针对跨两个表的联合查询，也做了摸底测试，联合查询耗时与单表查询没有明显区别。但是，如果应用场景发生变化，比如检索时长较长，经常需要联合查询或者需要联合多张表进行查询，MySQL 分表方案可能需要随之进行调整。

联合查询的 SQL 语句如下所示：

```
select Time from Table1
    where Time>=?5 and Time<=?6 and RA>=?1 and RA<=?2 and DEC>=?3 and DEC<=?4
    union
select Time from Table2
    where Time>=?5 and Time<=?6 and RA>=?1 and RA<=?2 and DEC>=?3 and DEC<=?4
```

4 面向应用的数据检索

本文方法能够支持以时间、空间为条件对特征参数进行检索。根据时间和空间组合数据检索条件，数据检索需求可分为时间点、时间范围、空间点、空间范围、时间点+空间点、时间点+空间范围、时间范围+空间点、时间范围+空间范围共 8 种情况。

当检索请求仅包含时间信息时，根据时间点或者时间范围对待检索参数表发起多个并行检索任务，针对以时间点为检索条件的请求，调用 HBase get 方法（根据唯一健值查询）对参数表进行检索；针对以时间段为检索条件的请求，调用 HBase scan 方法（根据唯一键值的起止范围查询）对参数表进行检索；完成检索后，对多个任务的检索结果进行合并。

当检索请求包含空间信息时，首先检索时空索引表，获取符合条件的时间信息；然后再根据时间信息检索参数表。

以检索赤经 RA 范围在（r1，r3）、赤纬 DEC 范围在（d1，d2）、观测时间 Time 范围在（t1，t100）且检索参数为 A、B、C、W 为例，检索过程如图 2 所示。

（1）以“r1<RA<r3 and d1<DEC<d2 and t1<Time<t100”为条件对时空索引表进行检索，得到检索结果①；

（2）以“t1<Time<t100”为条件，检索参数为 A、B、C、W，由于参数 W 与参数 A、B、C 存储在不同表，需生成 ABC 参数表检索任务和 W 参数表检索任务，两条检索并发执行；

（3）以“t1<Time<t100”为条件，首先查 ABC 参数表的分区索引，然后对符合条件的分区表同时进行检索，得到检索结果②、③；

（4）以“t1<Time<t100”为条件，首先查 W 参数表的分区索引，然后对符合条件的分区表同时进行检索，得到检索结果④、⑤；

（5）合并检索结果②、③、④、⑤，最终得到检索结果⑥。

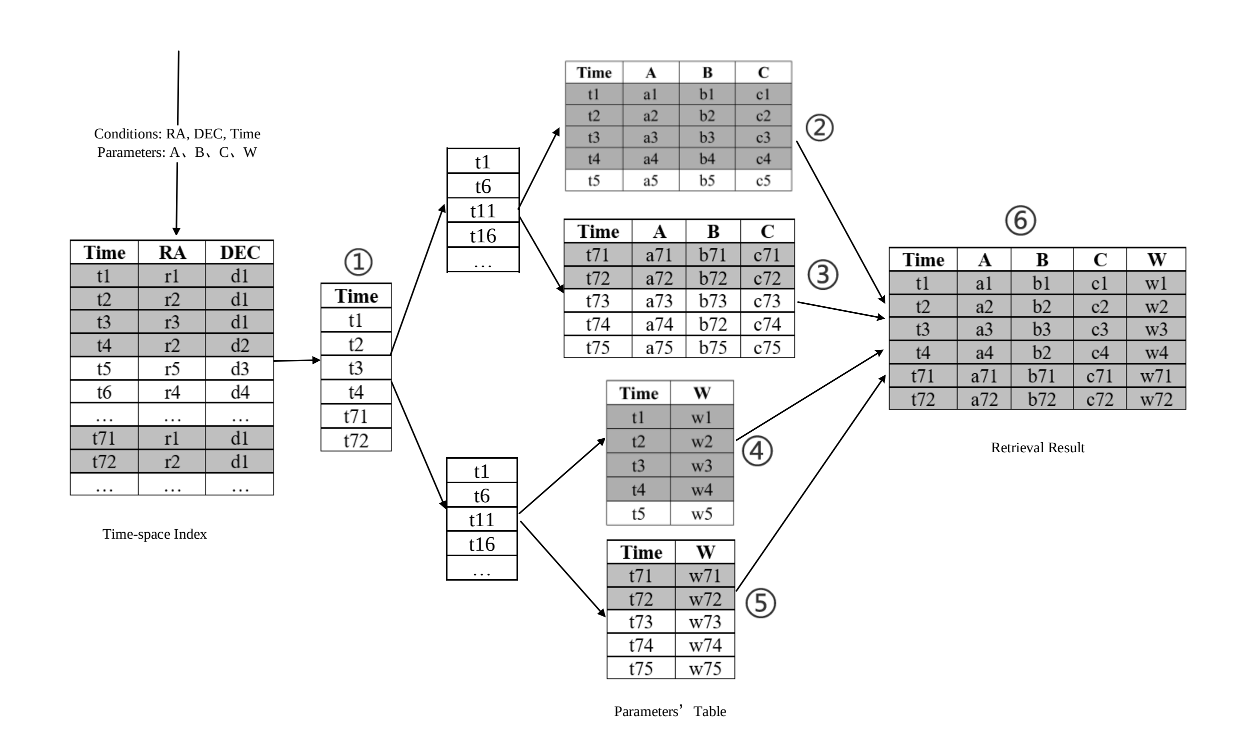


图 2 数据检索过程

Fig.2 the process of data retrieval

5 系统验证与分析

5.1 实验设计

按照本文方法，基于 3 台 4 核 CPU、32G 内存的虚拟服务器，搭建了一套 HBase 集群+MySQL 的测试系统（以下简称改进系统 HeteroDB）。测试数据为硬 X 卫星 2021 年 9 月 1 日零点至 2021 年 10 月 31 日零点，记录数量约为 500 万条。对照系统采用 MySQL 数据库，以库表形式存储了硬 X 卫星同时段探测数据（以下简称对照系统 MySQL）。

查询检索服务是数据组织和管理的核心，因此，本文基于数据检索效率对两种系统进行对比验证。检索条件选取了按照时间范围检索某类参数、按照时间范围检索多类参数、按照时间和空间范围联合检索某类参数共 3 类典型场景，检索条件涵盖时间、空间，检索参数涵盖单个、多个。针对上述场景，通过设置相同的检索条件，对比改进系统、对照系统的检索效率。为避免虚拟服务器资源不稳定带来的影响以及其他偶然误差，本文所有检索分两个时段进行，记录值为 10 次检索耗时的平均值。

实验 1 针对两种系统在时间检索场景下的响应速度进行对比。实验 2 针对改进系统对时间+空间检索场景下的响应速度进行测试。

进一步设计实验 3，针对本文方法在数据量增加情况下的可扩展性进行验证。实验 3 中，将 HBase 数据库的记录数量由 500 万条逐步扩展到 8000 万条，以相同的检索条件来测试数据检索耗时与数据规模之间的关系。

5.2 实验结果及分析

（1）实验 1：时间检索实验及结果分析

实验 1 设置为库表记录数量为 500 万条的情况下，以时间范围为检索条件获取指定参数，其中时间跨度为 1 小时、2 小时、3 小时、4 小时，检索参数为 1 个参数、3 个参数。设置相同的检索条件，分别在改进系统、对照系统中进行检索。

测试结果（

表 1）显示，在检索时间跨度较小、检索单个参数的场景下（场景 1-1），对照系统检索耗时与改进系统相当；但随着检索时间跨度的增长、参数的增多，对照系统耗时增长明显，在场景 2-4 中，前者耗时是后者的 80 多倍，显著超过改进系统。

这是由于 MySQL 数据库平衡二叉树索引机制，在检索过程中需要多次查找，随着检索时长增大、检

索参数增多，其查找次数增多，导致检索效率呈现成倍下降的趋势。改进系统基于 HBase 数据库以字典序排序方式存储行键的机制，采用时间作为行键，并且使用参数分表和时间分区存储、并行查询的方式，提高了检索效率，因此在面向时间检索场景中获得了明显的改进效果。

（2）实验 2：时空联合检索实验及结果分析

实验 2 检索条件设置为库表记录数量为 500 万条的情况下，按照时间范围、赤经范围、赤纬范围的联合条件检索单个参数，对比两种方法对时空联合检索请求的响应速度。

测试结果（表 2）显示，本文方法能够较好地支持以时间范围、空间范围联合对数据进行检索。场景 3-1（时间跨度 1 小时、赤经赤纬跨度 10 度）中，改进系统检索耗时为 26.3ms；同样检索条件下，对照系统检索耗时与改进系统基本相当。但随着时间范围、空间范围的扩大，对照系统检索耗时远超过改进系统。场景 3-4（时间跨度 4 小时、赤经赤纬 10 度）中，对照系统检索耗时约为改进系统的 5.3 倍。

（3）实验 3：扩展性实验与结果分析

为了验证改进系统在不同数据规模下的检索性能，将测试数据逐步扩展到 1000 万、2000 万、4000 万、8000 万，采用相同的检索场景来测试系统检索效率。测试结果（

表 3）表明，在测试数据量范围内，随着数据量的增大，改进系统对于时间检索、时空联合检索场景的检索效率保持基本稳定。

这是由于改进系统采用参数表按时间分区的存储方式，支持对符合检索条件的多个分区同时进行检索。因此，随着数据量的增大，时间分区也会增多，多分区并行检索的机制使得检索效率基本保持稳定。随着数据量的进一步增大，在时间分区数量过多、服务器资源不足的情况下，一定会出现检索效率逐渐降低的趋势，这时可以通过增加 HBase 分布式数据库的节点数量来保证数据检索效率。

表 1 时间检索效率对比

Table1 Comparison of efficiency under time retrieval conditions

No.	Retrieval Conditions	The Average Retrieval time of HeteroDB/ms	The Average Retrieval time of MySQL/ms
1-1	Time span: 1hour, Parameter number: 1	19.3	16.9
1-2	Time span: 2hours, Parameter number: 1	25.6	823.6
1-3	Time span: 3hours, Parameter number: 1	39.6	1935.3
1-4	Time span: 4hours, Parameter number: 1	50.2	2702.5
2-1	Time span: 1hour, Parameter number: 3	28.7	1065.7
2-2	Time span: 2hours, Parameter number: 3	49.5	3469.4
2-3	Time span: 3hours, Parameter number: 3	62.4	5653.2
2-4	Time span: 4hours, Parameter number: 3	97.6	7780.7

表 2 时空联合检索效率对比

Table2 Comparison of efficiency under time and space retrieval conditions

No.	Retrieval Conditions	The Average Retrieval time of HeteroDB/ms	The Average Retrieval time of MySQL/ms
3-1	Time span: 1hour, RA span and DEC span are both 10°, Parameter number: 1	26.3	29.6
3-2	Time span: 2hours, RA span and DEC span are both 20°, Parameter number: 1	221.3	454.7
3-3	Time span: 3hours, RA span and DEC span are both 30°, Parameter number: 1	393.7	1487.4
3-4	Time span: 4hours, RA span and DEC span are both 40°, Parameter number: 1	517.4	2736.7

表 3 不同数据规模下检索效率对比

Table3 Comparison of efficiency under different data sizes

No.	The Average Retrieval time of HeteroDB /ms（5millions）	The Average Retrieval time of HeteroDB /ms（10millions）	The Average Retrieval time of HeteroDB /ms（20millions）	The Average Retrieval time of HeteroDB /ms（40millions）	The Average Retrieval time of HeteroDB /ms（80millions）
1-1	19.3	20.5	20.8	21.6	21.2
1-2	25.6	25.3	24.5	26.3	27.6
1-3	39.6	39.4	40.2	42.9	46.5
1-4	50.2	49.8	53.8	55.6	54.4
3-1	26.3	26.1	28.6	28.7	29.2
3-2	221.3	230.5	221.8	222.9	231.8
3-3	393.7	378.2	390.1	403.2	404.3
3-4	517.4	513.1	515.1	521.9	525.2

6 结论

针对按照时间、空间范围快速获取空间天文卫星数据指定参数的需求，本文提出一种对海量数据进行高效组织的方法。区别于现有的以数据文件为粒度的管理方式，本方法通过建立观测时间、空间位置与特征参数的关联关系，将数据文件中的各类参数纳入到统一时空框架下；同时采用 HBase 分布式数据库存储各类参数，采用 MySQL 关系型数据库存储时空索引，以细粒度方式进行数据的组织管理。本方法所做改进包括以下几点：

（1）将卫星数据文件格式解析、参数抽取、物理量转换等过程从传统的数据获取过程中剥离出来，简化数据获取环节；

（2）构建分布式数据库集群，为参数独立建表，并且按照时间范围对参数表分区，将参数存储在分区表中，支持按时间并行检索，提高检索效率；

（3）采用关系型数据库分表的方式，存储空间天文卫星时空索引数据，支持从时间、空间两个维度检索数据；

（4）基于分布式数据库，对于观测时间、参数种类增加带来的数据量增长具备良好的可扩展性，能够适应数据持续增长的存储要求。

仿真结果表明，本文方法能够显著提高数据检索效率，满足实时获取空间天文卫星数据的应用需要。

参考文献

[1]. 熊森林等，空间科学数据产品组织模型[J]. 农业大数据学报, 2019, 1(4): 30-36.
Senlin Xiong, et al. Framework for Space Science Data Organization[J]. Journal of Agricultural Big Data, 2019, 1(4): 30-36.

[2]. Definition of the Flexible Image Transport System(FITS) , version 2.1b , December 9, 2005. http://fits.gsfc.nasa.gov/standard21b.html[1]

[3]. 杨晓艳等，FITS 变长数组在暗物质卫星数据存储中的应用研究[J]，天文研究与技术，2018(02): 176-180
Yang Xiaoyan, et al. Application Research of FITS Variable-Length Arrays in DAMPE Data Storage [J]. Astronomical Research & Technology, 2018(02): 176-180

[4]. Rademakers F, Brun R. ROOT: An object-oriented data analysis framework [J]. Nuclear Instruments & Methods in Physics Research, 1998, 389(1/2): 81-86

[5]. 崔辰州等，FITS 数据文件的检索和访问[J]，天文研究与技术，2008(02): 116-123
Cui Chengzhou, et al. Search and Location of FITS data files [J]. Astronomical Research & Technology, 2008(02): 116-123

[6]. 马福利等，GECAM 卫星快速预处理流程设计与实现[J]，天文研究与技术，https://doi.org/10.14005/j.cnki.issn1672-7673.20210611.001
Fuli Ma, et al. Design and Implementation of GECAM Preprocessing pipeline[J] . Astronomical Research &

Technology, https://doi.org/10.14005/j.cnki.issn1672-7673.20210611.001
[7]. 张嘉等. 大规模空间矢量数据分布式存储与计算优化[J]，计算机系统应用，2020,29(12):251–256
Zhang Jia, et al. Storage and Computing Optimization of Large Scale Distributed Spatial Vector Data[J]. Computer Systems & Applications,2020,29(12):251−256
[8]. 康栋贺，邹自明等.支持时空耦合计算的 HTM-ST 日地空间系统数据组织模型[J].地球信息科学学报,2017,19(6):735-743.
Kang D H, Zou Z M, Hu X Y, et al. 2017. HTM-ST: A data model supporting spatio-temporal coupled computation for solar-terrestrial system. Journal of Geo-information Science, 19(6):735-743.] DOI:10.3724/SP.J.1047.2017.00735
[9]. 葛微等. HiBase：一种基于分层式索引的高效 HBase 查询技术与系统[J]，计算机学报，2016，39（1）：140-153
Ge Wei, et al. HiBase: A Hierarchical Indexing Mechanism and System for Efficient HBase Query[J]. Chinese Journal of Computers. 2016,39(1):140-153
[10]. 袁茂林等. HBase 的高效时空分类索引[J]，小型微型计算机系统，2017，38（6）：1231-1236
Yuan Maolin, et al. Efficient Spatio-temporal Classification Index Under HBase[J]. Journal of Chinese Computer Systems. Vol. 38, No.6, 2017

Data organization method of space astronomical satellite based on Heterogeneous Database

YANG Xiaoyan[1,2] SUN Xiaojuan[1,2,3] SHI Tao[1,2] LIU Zhiqi[1] TONG Jizhou[4]
(1 Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China; 2 Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China; 3 Chinese Academy of Sciences University, Beijing 100149, China; 4 National Space Science Center, Chinese Academy of Sciences, Beijing 100149, China)

Abstract: With the increase of the space astronomical satellite data volume, data applications have gradually played an important role. Data applications such as research on space astronomical targets, payload status monitoring, spatial target analysis, detection plan assistance formulation, detection process visualization, etc., all require the processing and analysis of multi-source massive data such as detection target information, satellite platform parameters, and payload parameters. In order to realize the rapid extraction of characteristic data, this paper proposed a new method of data organization for space detection satellites. The method extracted massive characteristic parameters from data files, and established the association of observation time, spatial location and characteristic parameters, so as to realize multi-source data organization under a unified time and space frame. Then, Using the heterogeneous storage method of the combination of SQL database and No-SQL database, a characteristic parameter storage management system of 10 billion or even 100 billion was designed. The method in this paper was applied to the space science satellite big data application platform system. The experimental results of using HXMT(hard X-ray modulation telescope) satellite data showed that the system can better meet the requirements of obtaining data according to time and space conditions, compared with relational database data organization method, and the data retrieval efficiency under the same query mode was significantly improved; and with the increase of data storage capacity, the system has stable expansion capabilities.

Key words: Space astronomical satellite data; time-space index; data organization; No-SQL database